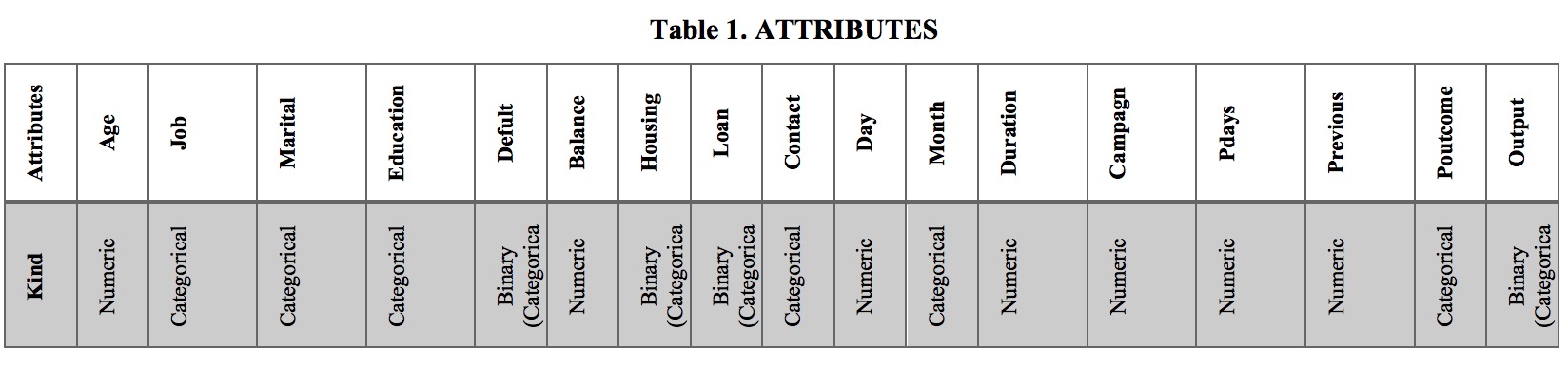
Apply Bayesian logistic regression in bank marketing prediction

STAT3016 final report by Lin Bai u5461367

**Introduction:**

The data analysed in this report is related with direct marketing campaigns of a Portuguese banking institution. The bank marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (‘yes’) or not (‘no’) subscribed. The data set consists 16 independent variables and 1 response variable as shown in the table below



The independent variables include 7 numeric variables and 9 categorical variables. The response variable ‘y’ indicates if the client subscribed a term deposit. We want to know which of these variables are important predictors that effects the probability of a client to subscribe the term deposit.

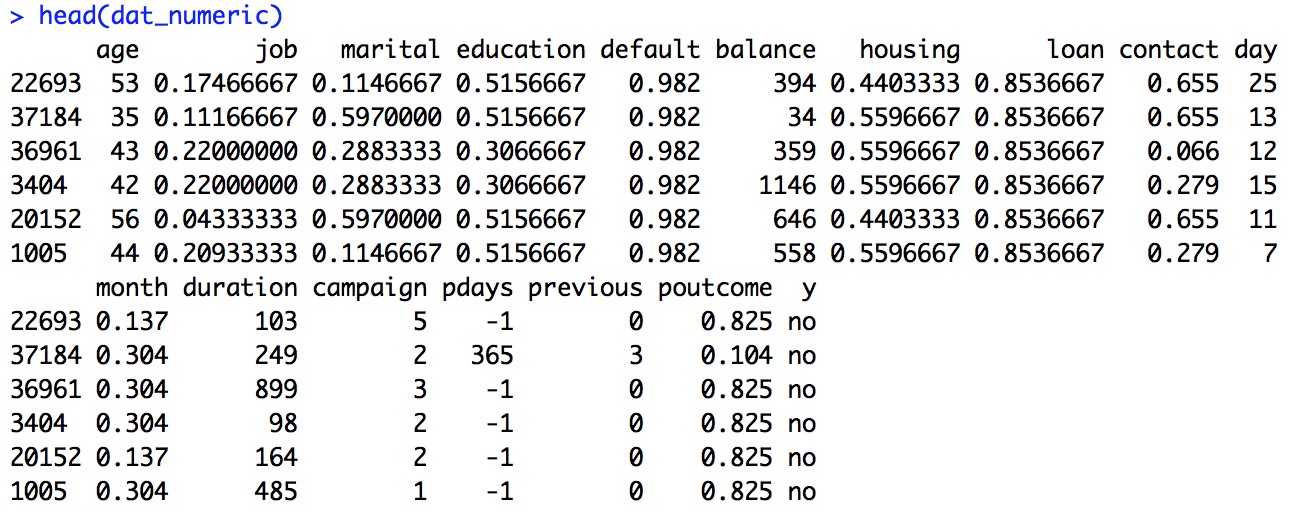
**Methodology**

Since the response variable is binary, i.e., ‘yes’ or ‘no’, we can formulate a binary classification problem using the data set. Binary classification problems can commonly be solved by a variety of models, among which logistic regression is quite popular because of its simplicity. Specifically, I would build up a logistic regression model under the Bayesian framework.

1. **Data preparation**

The data contains 45211 observations which are all complete, i.e., without missing values. Thus, there is no need to perform any missing value imputation. For convenience, we select 3000 of them as our sample data.

For the 9 categorical variables, I would encode them with the corresponding log-odd-ratio(log of the odds-ratios, which is used to qualify how strongly the presence or absence of A is associated with the presence or absence of property B) in order to transform them to numeric variables:



Before fit the data with a regression model, I would also scale the variables to make each variable centored and normalized.

1. **Model Specification**

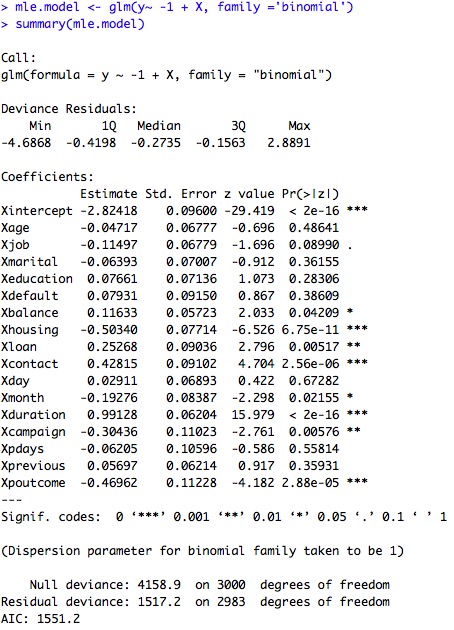
I use xi,j to denote variable j for sample i, and use Yi to denote its response. Consider a logistic regression model of the form Pr(Yi =1|**xi**,γ,β)=exp(θi)/(1+exp(θi)) where

θi = β0 + β1γ1xi,1 + ... + β16γ16xi,16

* In the model, γi is either 0 or 1 to indicate if the variable is selected or not. I assume the
* distribution of each γi is independent of each other and following a Bernoulli
* distribution. As for βj, I assume each one has independent normal prior distribution. We
* need to obtain the posterior distributions for **β** and **γ**.

1. **MCMC algorithm to approximate the posterior distribution**

In this section, I would implement the Monte Carlo Markov chain to approximate the posterior distribution of both β and γ. Specifically, the MCMC algorithm used in this study is Metropolis-Hastings, a generic method of approximating the posterior distribution corresponding to any combination of prior distribution and sampling model.  To start the MCMC approximation, I would first fit the logistic regression by maximum likelihood estimation in order to set proper parameters for the prior distributions of **β**.



Steps of the algorithm:

For **β**, sample β∗ ∼ J(**β**| **β**(s)) (where J(**β**∗| **β**(s)) is a symmetric proposal distribution), in this case, the proposal distribution is multivariate normal(βn, Σn). βn is the coef(mle.model), Σn is the summary(mle.model)$cov.unscaled in the mle model.

Then calculate the acceptance ratio, we have:

##lpy.c==current log-likelihood

lpy.c<-sum(dbinom(y,1,ilogit(X[,gamma==1,drop=FALSE]%\*%beta[gamma==1]),log=T))

log(r)=log(y|x,beta\*)-log(y|x,beta)  +log( p(beta\*))-log( p(beta))

>lpy\_beta.p<-sum(dbinom(y,1,ilogit(X[,gamma==1,drop=FALSE]%\*%beta.p[gamma==1]),log=T))

>lhr.beta<-lpy\_beta.p-lpy.c+sum(dnorm(beta.p,pmn.beta,psd.beta,log=T))-sum(dnorm(beta,pmn.beta,psd.beta,log=T))

After getting the acceptance ratio, set β(s+1) to β∗ or β(s) with probability min(1,r) and

Max(0,1-r). If log(u) < log(r), setting β(s+1) = β∗ . Otherwise set β(s+1) = β(s) . In r we write:

if(log(runif(1))<lhr.beta) {beta<-beta.p; acs\_beta<-acs\_beta+1}

For **γ,** the steps are similar. The proposal distribution of γis Bernoulli(oj/(1+oj)).

o\_j=p(y|x,gamma\_{-j},gamma\_j=1)/p(y|x,gamma\_{-j},gamma\_j=0), r\_j=log(o\_j). gamma\_j ~Bern(o\_j/(1+o\_j))=Bern(1/(1+exp(-r\_j))) ##o\_j is the conditional odds that gamma[j]=1

gamma\_p<-gamma ; gamma\_p[j]<-1-gamma\_p[j] #lpy.p==proposal loglikelihood

lpy.p<-sum(dbinom(y,1,ilogit(X[,gamma\_p==1,drop=FALSE]%\*%

beta[gamma\_p==1,drop=FALSE]),log=T))

lhr<-(lpy.p-lpy.c)\*(-1)^(gamma\_p[j]==0)

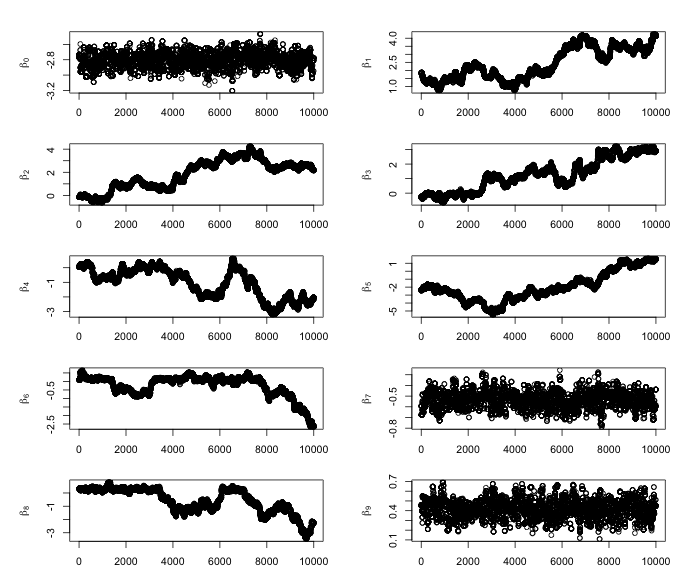
gamma[j]<-rbinom(1,1,1/(1+exp(-lhr)))

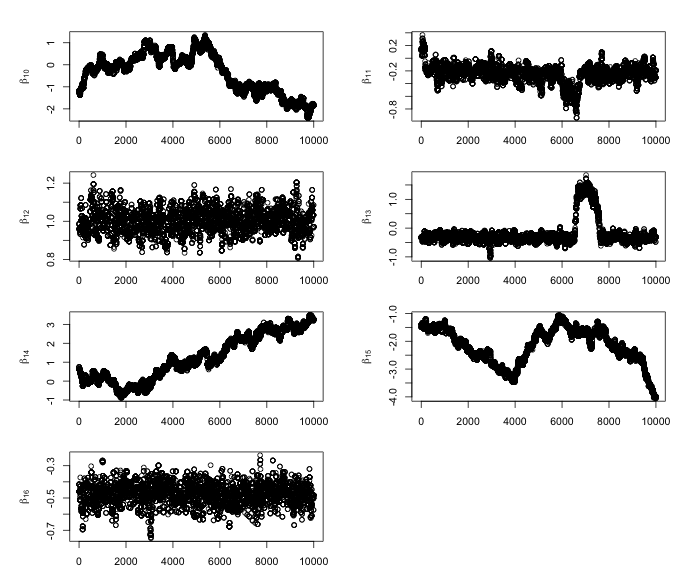
if(gamma[j]==gamma\_p[j]) {lpy.c<-lpy.p} ##which means if (log(runif(1))<lhr) {gamma[j]<-gamma\_p[j] ; lpy.c<-lpy.p}

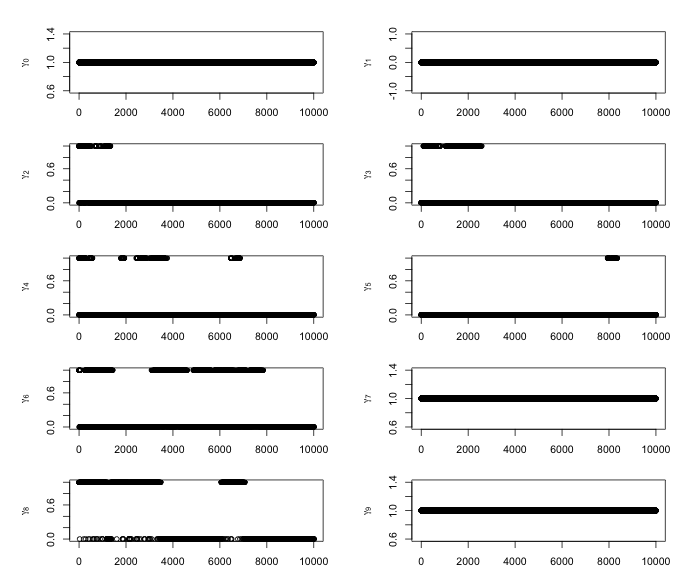
1. **Result Analysis**

Based on the MCMC results, I plot the traceplots for βj(**s**), γj(**s**) and γj(**s**)βj(**s**)

Figure1: Trace plots of βj(**s**)



Figure2: Trace plots of γj(**s**)



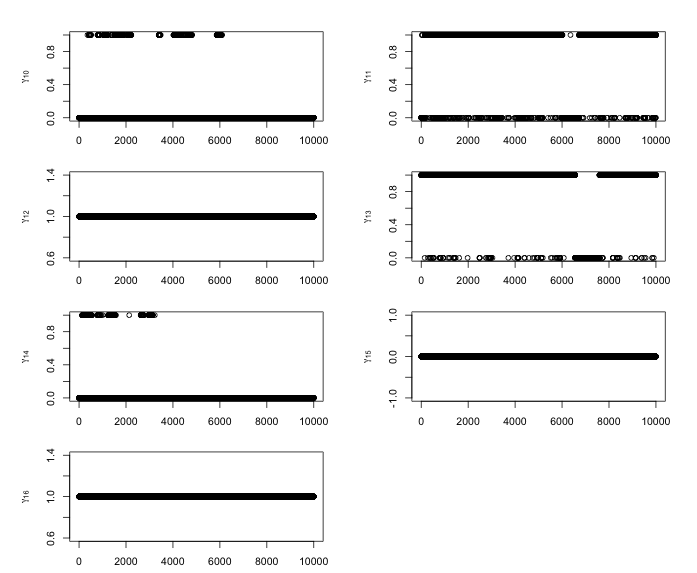
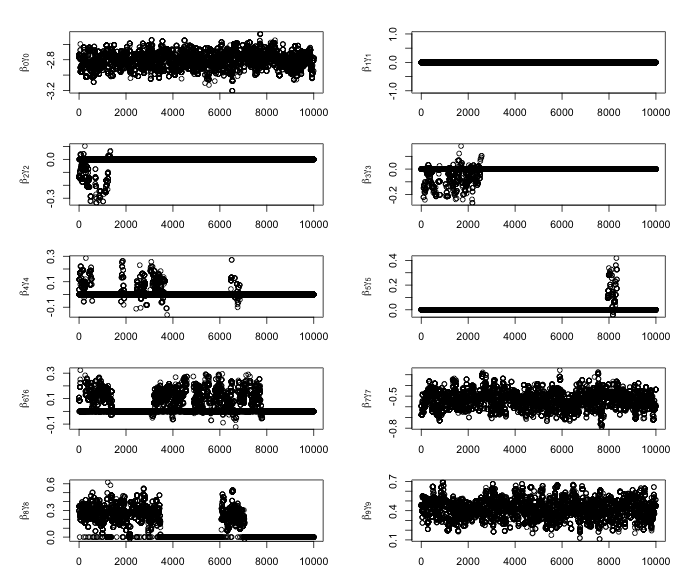


Figure3: Trace plot of γj(**s**)βj(**s**)



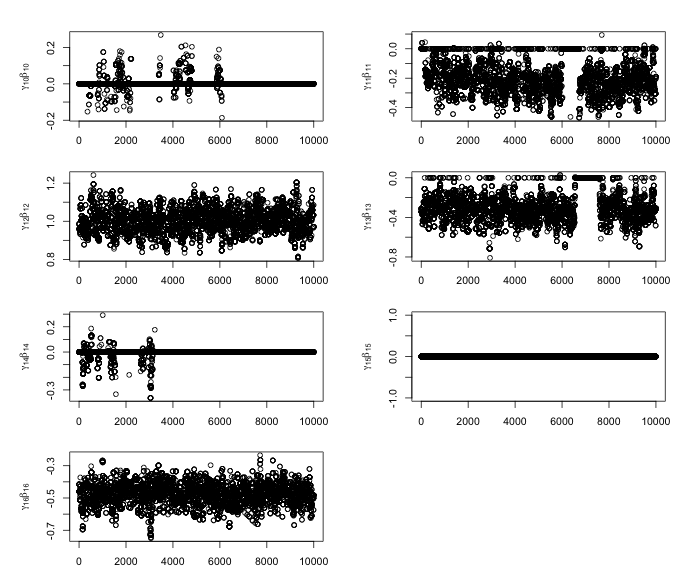
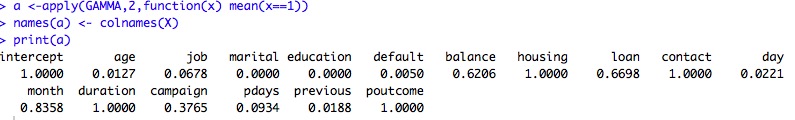


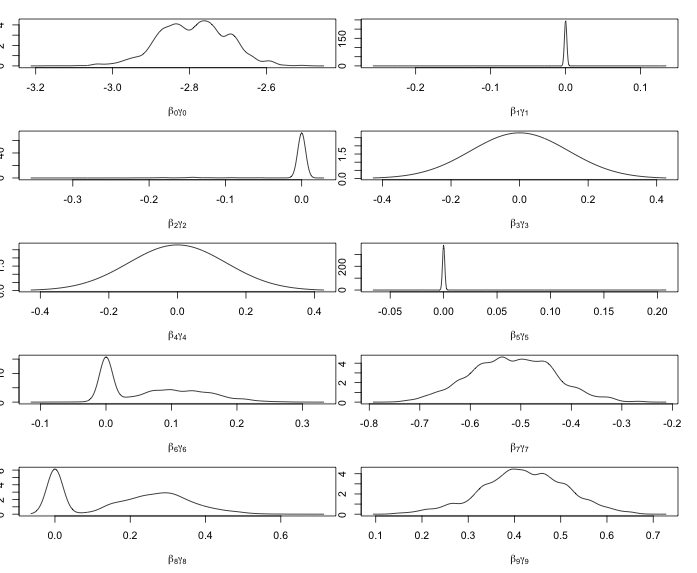
Figure1 only shows stationarity of the MCMC chains for β0 , β7 , β9 β12and β16, which is reflective of the predominant draws of γ1(s) =γ2(s) = γ3(s) = γ4(s)= γ5(s) = γ6(s) = γ8(s) = γ10(s) = γ11(s) = γ13(s) = γ14(s) = γ15(s) = 0. Occasionally we accept draws of the above gammas to be 1, as displayed in Figures 2 and 3, which results in the irregular peaks and troughs in the MCMC chains for corresponding β in Figure 1. That is, our results indicate that the following variables are not important predictors of the response of the client:

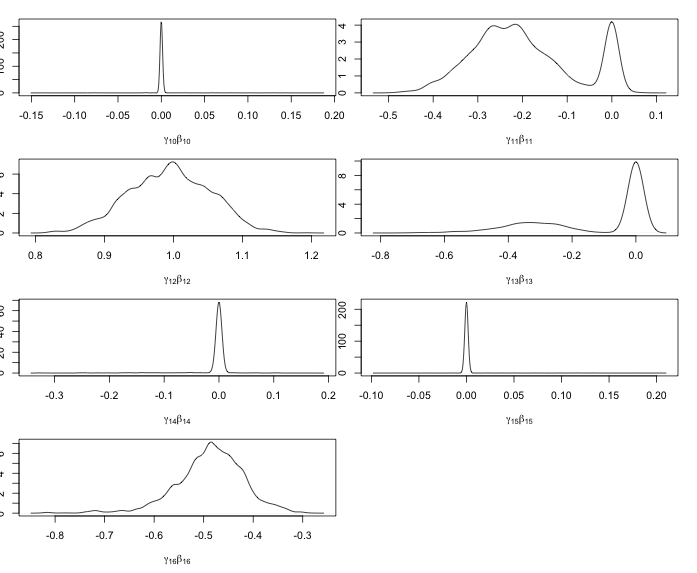
“age”, “job”, “material”, “education”, “default”, “balance”, “loan”, “day”, “month”, “campaign”, “pdays”, “previous”



The values for Pr(γj =1|x,y) are displayed above and indicate that except for those important variables, "age", "job", "marital", "education", "default", "day", "pdays" and "previous" have weak association with the outcome. "balance", “loan”, “month” and “campaign” have relatively strong association with the outcome.

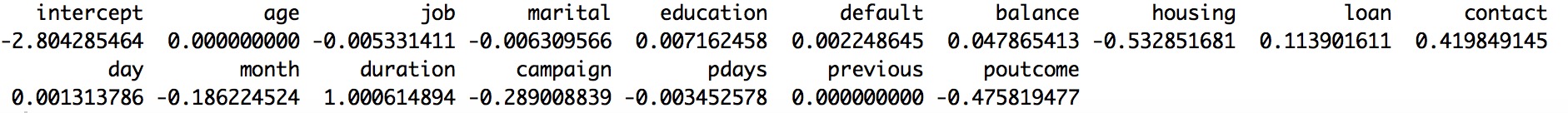
Figure4: Posterior density plot and posterior mean estimate for βjγj



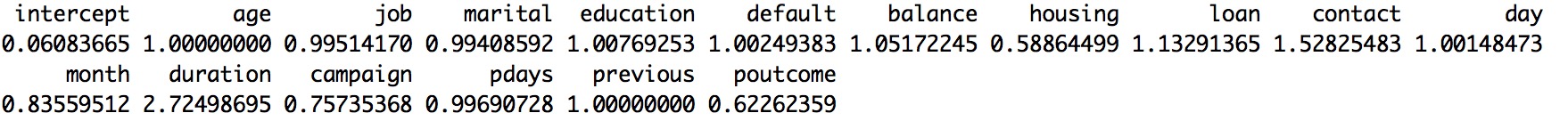


From figure 4 we can see that for the not important predictors, their density mostly gather among 0 and the peak of their density plots are at 0. The posterior density of β6γ6, β8γ8 β13γ13 are binomial, which is reflected in the value of P r(γ1 = 1|x, y) (0.6206, 0.6698, 0.3765). Those values are not approximate to 0 or 1 like other variables.

The posterior mean estimates of βjγj :



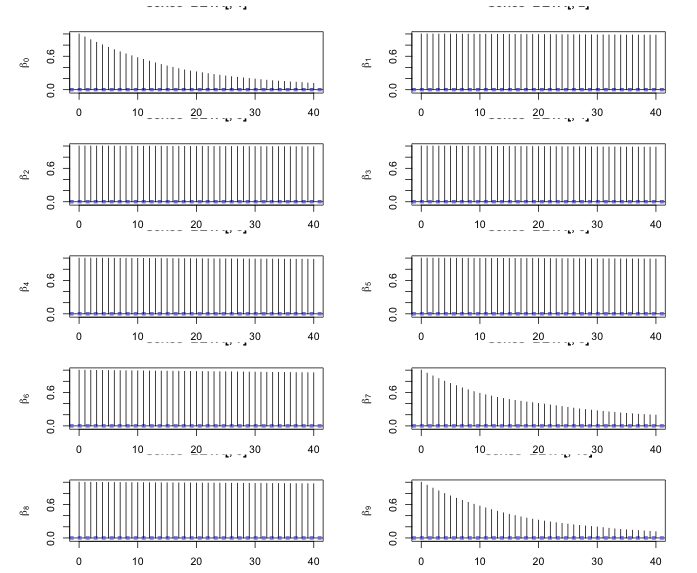
And the posterior mean estimates of exp(βjγj) :



1. **Model checking**

We use the auto-correlation to see whether the MCMC sample is clumpy, to find whether the sample is representative or not.

Figure 5: Autocorrelation plots for βj(s)



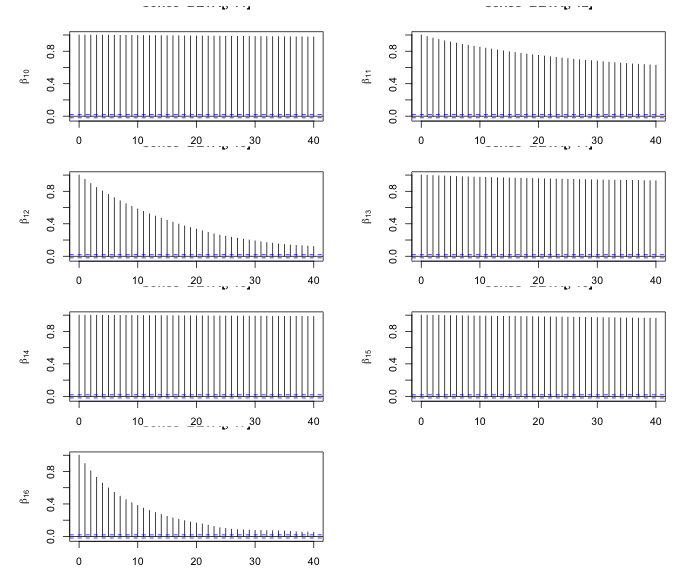
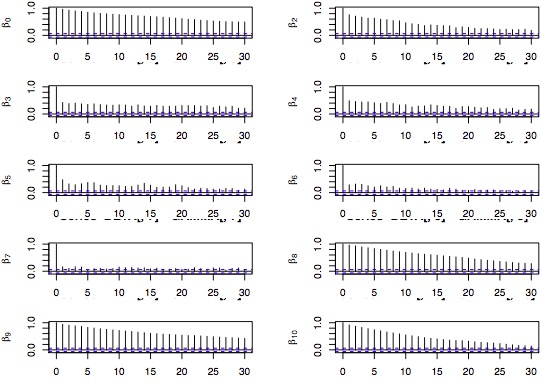
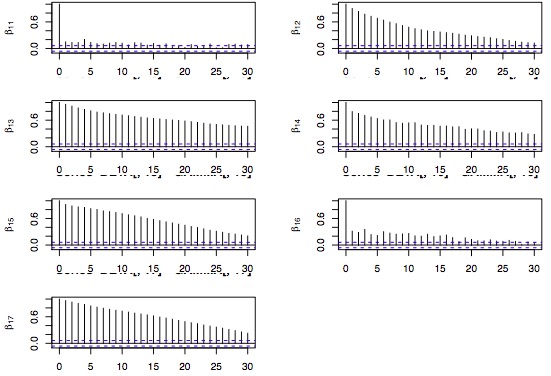


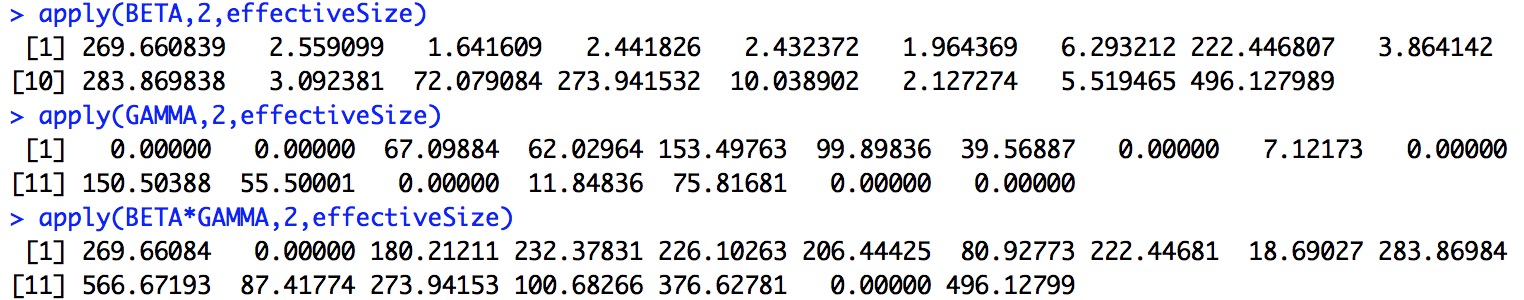
Figure6: Autocorrelation for γjβj





However, for the important predictors, the autocorrelation plots show high autocorrelations to be concerned about, so the current estimates of Pr(γj = 1|x,y) are not reliable. The sequences should be thinned and/or the chains run for a longer number of iterations and estimates for Pr(γj=1|x,y) recalculated. Also note that the estimates in parts (b) and (c) are Bayesian model averaged estimates, because we average over all iterations where the variables are not always active in the logistic model, depending on whether γj(s) = 1 or 0.

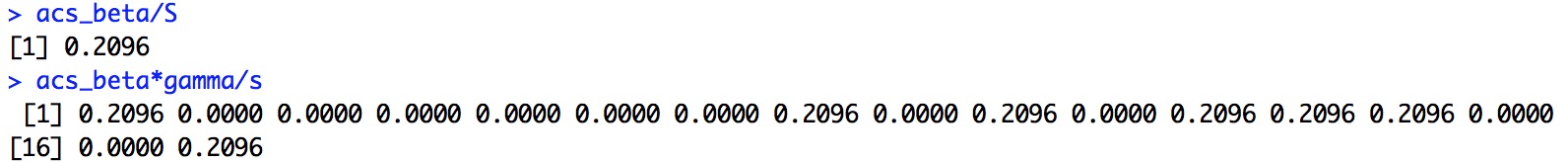
Effective size



From the effectiveSize outcome we can see that, for the important predictors (the 0th, 7th, 9th, 12th , 16th ), the effectiveSize is relatively larger. But overall they are still small comparing to 10000 times of algorithm.

Acceptance rate

The acceptance rate for βj and βjγj of the predictors is 0.2096, which is a little bit low but reasonable.



**Conclusion**

We can see from the posterior mean estimates for exp(βjγj) and the analysis above that the predictors ”housing”, “contact”, “duration” and “poutcome” are important predictors of the response. Except for “duration”, the other predictors have negative correlation with the outcome. The estimated change in odds of subscribe ranges from -41.2% to 172.49%. If other resources are available, we can make use of the whole 45211 observations and further resources to explore whether there are other factors which may affect the outcome.

**Self-criticism**

There are 3 main limitations of my analysis:

* For convenience(mainly the time it spent on running algorithm), I selected 3000 of the total 45,211 observations. This may reduce the accuracy of the predictions.
* The autocorrelation plots show high autocorrelations to be concerned about. So the current posterior estimates of γ are not reliable. One option is to thin the MCMC process.
* The effective size is relatively small comparing to the size 10000 of the iteration. The acceptance rate of 21% is also reasonable but relatively small. One option to enlarge the effective size and the acceptance rate is to choose another tunning parameter.